Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

dbEmpLikeGOF: An R Package for Nonparametric Likelihood Ratio Tests for Goodness-of-Fit and Two Sample Comparisons Based on Sample Entropy*

Jeffrey C. Miecznikowski
University at Buffalo

Albert Vexler
University at Buffalo

Lori Shepherd
Roswell Park Cancer Institute

Abstract

We introduce and examine **dbEmpLikeGOF**, an R language for performing goodness-of-fit tests based on sample entropy. This package also performs the two sample distribution comparison test. For a given vector of data observations, the provided function `dbEmpLikeGOF` tests the data for the proposed null distributions, or tests for distribution equality between two vectors of observations. The proposed methods represent a distribution-free density-based empirical likelihood technique applied to nonparametric testing. The proposed procedure performs exact and very efficient p values for each test statistic obtained from a Monte-Carlo (MC) resampling scheme. Note by using an MC scheme, we are assured exact level α tests that approximate nonparametrically most powerful Neyman-Pearson decision rules. Although these entropy based tests are known in the theoretical literature to be very efficient, they have not been well addressed in statistical software. This article briefly presents the proposed tests and introduces the package, with applications to real data. We apply the methods to produce a novel analysis of a recently published dataset related to coronary heart disease.

Keywords: empirical likelihood, likelihood ratio, goodness-of-fit, sample entropy, nonparametric tests, normality, two-sample comparisons, uniformity.

*This research is supported by the NIH grant 1R03DE020851 - 01A1 (the National Institute of Dental and Craniofacial Research)

1. Introduction

1.1. Empirical likelihood

Empirical likelihood (EL) allows researchers the benefit of employing powerful likelihood methods (maximizing likelihoods) without having to choose a parametric family for the data. A thorough overview of empirical likelihood methods can be found in Owen (2001). The research in this area continues to grow while empirical likelihood methods are being extended to many statistical problems as in, for example, Vexler, Yu, Tian, and Liu 2010; Yu, Vexler, and Tian 2010.

In short, an outline of the EL approach can be presented as follows. Given independently identically distributed observations $X_1, \ldots, X_n$, the EL function has the form of $L_p = \Pi_{i=1}^n p_i$ where the components $p_i, i = 1, \ldots, n$ maximize the likelihood L_p (maximum likelihood estimation) provided that empirical constraints, based on $X_1, \ldots, X_n$ are in effect ($\sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i X_i = 0$, under the hypothesis $EX_1 = 0$). Computation of the EL's components $p_i, i = 1, \ldots, n$ used to be an exercise in Lagrange multipliers. This nonparametric approach is a product of the consideration of the 'distribution-functions'-based likelihood $\Pi_{i=1}^n F(X_i) - F(X_i-)$ over all distribution functions F where $F(X_i-)$ denotes the left hand limit of F at X_i.

The following extensions from these methods involve a density-based likelihood methodology for goodness-of-fit testing. The proposed extensions have been motivated by developing test statistics that approximate nonparametrically most powerful Neyman-Pearson test statistics based on likelihood ratios. A density-based EL methodology can be introduced utilizing the EL concept as in Vexler and Gurevich 2010a,b; Gurevich and Vexler 2011. Following the EL methodology, the likelihood function $L_f = \Pi_{i=1}^n f(X_i)$ where $f(\cdot)$ is a density function of X_i can be approximated by $\Pi_{i=1}^n f_i$, where values of f_i should maximize $\Pi_{i=1}^n f_i$ provided that an empirical constraint which corresponds to $\int f(u)du = 1$ under an underlying hypothesis is in effect. Outputs of the density based EL approach have a structure that utilize sample entropy (for example, Vexler and Gurevich 2010a). To date, density based EL tests have not been presented in R packages (R Development Core Team 2009) but are known to be very efficient in practice. Moreover, despite the fact that many theoretical articles have considered very powerful entropy-based tests, to our knowledge there does not exist software procedures to execute procedures based on sample entropy in practice.

1.2. Goodness-of-fit tests

Goodness-of-fit tests commonly arise when researchers are interested in checking whether the data come from an assumed parametric model. In certain situations, this question manifests to test whether two datasets come from the same parametric model. Commonly used goodness-of-fit tests include the Shapiro-Wilks (SW) test, Kolmogorov-Smirnov (KS) test, the Lilliefors (L) test, Wilcoxon rank sum test (WRS), the Cramér-von-Mises test, and the Anderson-Darling test (Darling 1957; Lilliefors 1967; Hollander, Wolfe, and Wolfe 1973; Royston 1991).

Recently several new goodness-of-fit tests have been developed using density based empirical likelihood methods. These powerful new tests offer exact level α tests with critical values that can be easily obtained via Monte-Carlo approaches.

1.3. EL ratio test for normality

The derivation of the EL ratio test for normality can be found in Vexler and Gurevich (2010b). To outline this method, we suppose that the data consist of n independent and identically distributed observations $X_1, \ldots, X_n$. Consider the problem of testing the composite hypothesis that a sample $X_1, \ldots, X_n$ is from a normal population. Notationally, the null hypothesis is

$$H_0 : X_1, \ldots, X_n \sim N(\mu, \sigma^2), \tag{1}$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with unknown mean μ and unknown standard deviation σ. Generally speaking when the density functions f_{H_1} and f_{H_0} corresponding to the null and alternative hypotheses, are completely known, the most powerful test statistic is the likelihood ratio:

$$\frac{\prod_{i=1}^{n} f_{H_1}(X_i)}{\prod_{i=1}^{n} f_{H_0}(X_i)} = \frac{\prod_{i=1}^{n} f_{H_1}(X_i)}{(2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2\right)}, \tag{2}$$

where, under the null hypothesis, $X_1, \ldots, X_n$ are normal with mean μ and variance σ^2. In the case of the unknown μ and σ^2, the maximum likelihood estimation applied to (2) changes the ratio to,

$$\frac{\prod_{i=1}^{n} f_{H_1}(X_i)}{(2\pi e s^2)^{-n/2}}, \tag{3}$$

where s represents the sample standard deviation.

Applying the maximum EL method to (3) forms the likelihood ratio test statistic

$$T_{mn} = (2\pi e s^2)^{n/2} \prod_{i=1}^{n} \frac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)}, \tag{4}$$

where m is assumed to be less than $n/2$. Using empirical likelihood modifications, the maximum EL method applied to (3), and following Vexler and Gurevich (2010b), to test the null hypothesis at (1) we can use the test statistic,

$$V_n = \min_{1 \le m < n^{1-\delta}} (2\pi e s^2)^{n/2} \prod_{i=1}^{n} \frac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)} \tag{5}$$

where $0 < \delta < 1$, s denotes the sample standard deviation, and $X_{(1)}, \ldots, X_{(n)}$ represent the order statistics corresponding to the sample $X_1, \ldots, X_n$. Note, here, $X_{(j)} = X_{(1)}$ if $j \le 1$ and $X_{(j)} = X_{(n)}$ if $j \ge n$.

We employ the following decision rule, we reject the null hypothesis if and only if

$$\log(V_n) > C, \tag{6}$$

where C is a test threshold and V_n is the test statistic defined in (5).

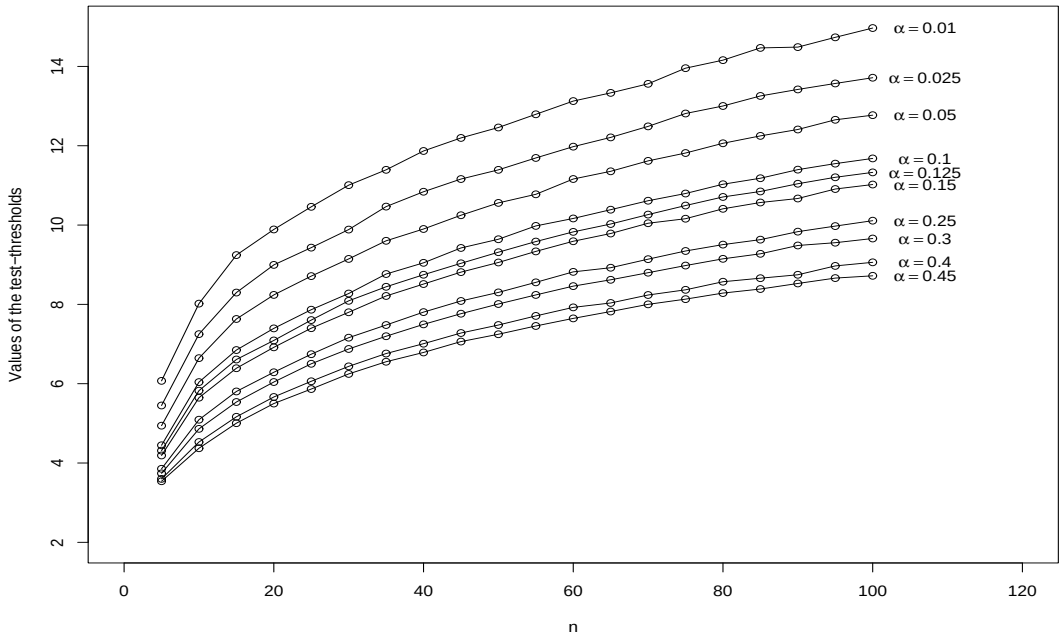


Figure 1: The curves display the value of the thresholds C_α for the test statistic $\log(V_n)$ with $\delta = 0.5$ corresponding to the significance (α) levels of $\alpha = 0.01, 0.025, 0.05, 0.125, 0.15, 0.25, 0.3, 0.4, 0.45$ that are plotted against the sample sizes $n = 5, 10, 15, \ldots, 100$.

Since

$$\sup_{\mu,\sigma} P_{H_0}\{\log(V_n) > C\} = P_{X_1,\ldots,X_n \sim N(0,1)}\{\log(V_n) > C\}, \tag{7}$$

the type I error of the test at (6) can be calculated exactly using a Monte Carlo approach. Type I error for the test in (6) refers to the probability of rejecting the null hypothesis in (1) when, in fact, the null hypothesis is true. Figure 1 displays the Monte-Carlo roots C_α of the equation $P_{X_1,\ldots,X_n \sim N(0,1)}\{\log(V_n) > C_\alpha\} = \alpha$ for different values of α and n. (For each value of α and n, the solutions were derived from 75,000 samples of size n.) The setting of $\delta = 0.5$ is motivated by the work presented in Vexler and Gurevich (2010b). In general, the choice of δ is not critical for these goodness-of-fit tests.

1.4. EL ratio test for uniformity

One can show that tests for uniformity correspond to general goodness-of-fit testing problems when the null hypothesis is based on completely known distribution functions. The full derivation of the EL ratio test for uniformity can be found in Vexler and Gurevich (2010b). We consider the test for the uniform distribution on the interval $[0, 1]$ ($Uni(0, 1)$), specifying the null distribution

$$H_0 : Y_1, \ldots, Y_n \sim Uni(0, 1) \tag{8}$$

versus the alternative that $Y_1, \ldots, Y_n$ are from a nonuniform distribution $F(y)$.

Before considering the hypothesis in (8), consider the problem of testing

$$H_0 : f = f_{H_0} \text{ vs } H_1 : f = f_{H_1}, \tag{9}$$

where, under the alternative hypothesis, f_{H_1} is completely unknown and under the null hypothesis $f_{H_0}(x) = f_{H_0}(x; \boldsymbol{\theta})$ is known up to the vector of parameters $\vec{\theta} = (\theta_1, \ldots, \theta_d)$, where $d \geq 1$ defines a dimension of the vector $\boldsymbol{\theta}$. In accordance with maximizing EL, for the test in (9) we obtain the statistic,

$$G_n = \min_{1 \leq m < n^{1-\delta}} \frac{\prod_{i=1}^{n} \frac{2m}{n(X_{(i+m)} - X_{(i-m)})}}{\prod_{i=1}^{n} f_{H_0}(X_i; \hat{\theta})}. \tag{10}$$

Applying the result in (10) to the specific hypothesis in (8) and using the outputs from Vexler and Gurevich (2010b), we suggest the following EL ratio test statistic

$$U_n = \min_{1 \leq m < n^{1-\delta}} \prod_{i=1}^{n} \frac{2m}{n\left(Y_{(i+m)} - Y_{(i-m)}\right)}, \tag{11}$$

where $0 < \delta < 1$ and $Y_{(1)}, \ldots, Y_{(n)}$ correspond to the order statistics from the sample $Y_1, \ldots, Y_n$. Note, $Y_{(j)} = Y_{(1)}$ if $j \leq 1$ and $Y_{(j)} = Y_{(n)}$ if $j \geq n$. The event

$$\log(U_n) > C \tag{12}$$

implies that H_0 is rejected, where C is a test threshold. The significance level of this test can be calculated according to the following equation,

$$P_{H_0}\left\{\log(U_n) > C_\alpha\right\} = P_{X_1, \ldots, X_n \sim Uni(0,1)}\left\{\log(U_n) > C_\alpha\right\} = \alpha. \tag{13}$$

Figure 2 shows the roots C_α of the equation in (13) for different values of α and n. (For each value of α and n, the solution is derived from 75,000 samples of size n).

Note, the test for uniformity in (12) will cover a generalized version of the goodness-of-fit problem when the distribution in H_0 is completely known. In other words, if we consider the random sample $X_1, \ldots, X_n$ from a population with a density function f and a finite variance we can test the hypotheses:

$$H_0 : F = F_{H_0} \text{ vs } H_1 : F = F_{H_1}, \tag{14}$$

where, under the alternative hypothesis, F_{H_1} is completely unknown, whereas under the null hypothesis, $F_{H_0}(x) = F_{H_0}(x; \boldsymbol{\theta})$ is known up to the vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. Note, $d \geq 1$ defines the dimension for $\boldsymbol{\theta}$. Although a strong assumption, by assuming that densities exist under the alternative, we are able to demonstrate asymptotic consistency of the proposed test statistic. By employing the probability integral transformation (Dodge 2006), if $X_1, \ldots, X_n \sim f_{H_0}$, with f_{H_0} completely known, then $Y_i = F_{H_0}^{-1}(X_i) \sim Uni(0, 1)$. Hence, the uniformity test in (12) can be employed on data $Y_1, \ldots, Y_n$ to test whether $X_1, \ldots, X_n$ conforms with density f_{H_0}.

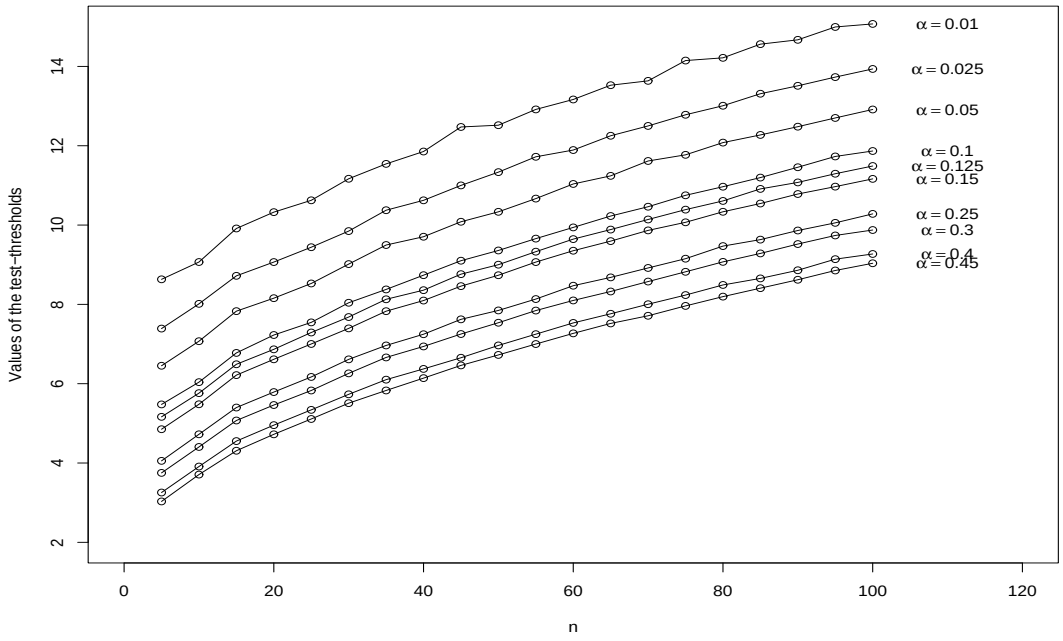


Figure 2: The curves display the value of the thresholds C_α for the test statistic $\log(U_n)$ with $\delta = 0.5$ corresponding to the significance (α) levels of $\alpha = 0.01, 0.025, 0.05, 0.125, 0.15, 0.25, 0.3, 0.4, 0.45$ that are plotted against the sample sizes $n = 5, 10, 15, \ldots, 100$.

1.5. EL ratio test for distribution equality

In this section we present the EL ratio test for examining if two datasets are from the same distribution. The complete derivation for this case can be found in Gurevich and Vexler (2011). In short, let $X_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$ denote independent observations in the first dataset and $X_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$ denote independent observations in another dataset. Under H_0 (equal distributions), we assume that both groups are identically distributed. That is, our null hypothesis is

$$H_0 : F_{X_1} = F_{X_2} \tag{15}$$

where F_{X_1} and F_{X_2} denote the cumulative density function (CDF) for the observations in X_1 and X_2, respectively.

To derive the test for (15), we consider that the likelihood ratio can be expressed as,

$$R = \frac{\prod_{i=1}^{n}\prod_{j=1}^{n_i} f_{X_i}(x_{i(j)})}{\prod_{i=1}^{n}\prod_{j=1}^{n_i} f_{X}(x_{i(j)})}. \tag{16}$$

Following the EL concept, we approximate the likelihoods and integrals and obtain the non parametric approximation to (16) as,

$$\tilde{R}_{m,v,n_1,n_2} = \prod_{i=1}^{n_1} \frac{2m}{n_1 \delta_{m1j}} \prod_{j=1}^{n_2} \frac{2v}{n_2 \delta_{v2j}}. \tag{17}$$

The proper selection of m and v in the current literature of entropy-based decision making recommends selecting values utilizing information regarding alternative distributions when sample sizes are finite. Ultimately using the work in Yu *et al.* (2010), we look at selecting m and v by minimizing $\tilde{R}$ over appropriate ranges. This suggests the following test statistic for the hypothesis in (15),

$$\tilde{R}_{n_1,n_2} = \min_{l_{n_1} \le m \le u_{n_1}} \prod_{j=1}^{n_1} \frac{2m}{n_1 \Delta_{m1j}} \min_{l_{n_2} \le v \le u_{n_2}} \prod_{j=1}^{n_2} \frac{2v}{n_2 \Delta_{v2j}}, \tag{18}$$
$$l_n = n^{0.5+\delta}, u_n = \min(n^{1-\delta}, n/2), \delta \in (0, 0.25).$$

The Δ_{mij} function is defined as:

$$\Delta_{mij} = \frac{1}{n_1 + n_2} \sum_{k-1}^{2} \sum_{i=1}^{n_i} \left(I(x_{kl} \le x_{i(j+m)}) - I(x_{kl} \le x_{i(j-m)}) \right), \tag{19}$$

where $I()$ denotes an indicator function that takes the value 1 if the condition in the parenthesis is satisfied and takes the value 0, otherwise. The $x_{i(j)}$ indicates the j-th order statistic for the group i. Note, here, $x_{i(j+m)} = x_{i(n_i)}$, if $j + m \ge n_i$ and $x_{i(j-m)} = x_{i(1)}$ if $j - m \le 1$.

The test rejects the null hypothesis for large values of $\log \tilde{R}_{n_1,n_2}$. Note that we define $\Delta_{lij} = 1/(n_1 + n_2)$ if $\Delta_{lij} = 0$.

Significance of level α can be determined since $I(X > Y) = I(F(X) > F(Y))$ for any distribution function F. Hence, the null distribution of $\tilde{R}_{n_1,n_2}$ is independent with respect to the form of the underlying distributions given H_0. Hence, we can tabulate universal critical values regardless of the null distribution of the X_{ij}'s.

Table 1 shows the critical values for the logarithm of $\tilde{R}_{n_1,n_2}$ for common sample sizes and significance levels. These critical values were obtained from deriving Monte Carlo roots of

$$P_{H_0}(\log(\tilde{R}_{n_1,n_2}) > C_\alpha) = \alpha$$

based on 75,000 repetitions of sampling $X_{1j} \sim N(0,1)$ and $X_{2j} \sim N(0,1)$.

In the following we present the structure and functioning of the package, with applications to real datasets.

2. What is package dbEmpLikeGOF

In summary, the **dbEmpLikeGOF** package provides a function `dbEmpLikeGOF` to be used for empirical likelihood based goodness-of-fit tests based on sample entropy. The function can also perform the two sample EL ratio test for the hypothesis in (15). The output of `dbEmpLikeGOF` analysis is an object containing the test statistic and the p value. Standard bootstrap options

n_1 \ n_2	α	10	15	20	25	30	35	40	45	50
10	0.01	10.0309	11.7135	12.6062	12.3904	13.4997	14.2343	14.9241	15.2228	15.2992
	0.03	9.2193	10.4764	11.4159	11.2406	12.2012	12.9932	13.455	13.6251	14.2996
	0.05	8.6969	10.1189	10.7509	10.6318	11.4996	12.4137	12.9247	13.031	13.7022
	0.1	7.9816	9.4214	9.9363	9.9797	10.7017	11.3997	12.0698	12.0546	12.6841
	0.3	6.8933	8.1363	8.693	8.8636	9.4652	10.0897	10.709	10.6893	11.2914
15	0.01	11.8648	12.458	13.2226	14.3837	14.3508	15.042	15.8725	15.9455	15.8574
	0.03	10.5947	11.4898	12.3084	12.8717	13.1676	13.9621	14.7072	14.8819	14.9636
	0.05	10.1105	10.9926	11.5624	12.1058	12.6275	13.4878	14.0286	14.2757	14.5255
	0.1	9.3739	10.3755	10.9992	11.282	11.8066	12.5398	13.1243	13.4726	13.6302
	0.3	8.1207	9.2095	9.9684	10.0254	10.6899	11.3412	11.9266	11.9816	12.564
20	0.01	12.1399	13.6004	13.7649	14.4306	15.2331	16.9505	16.315	16.1152	16.8527
	0.03	11.0597	12.1414	12.7128	13.2027	13.8217	14.705	15.0791	15.1986	15.8462
	0.05	10.5834	11.6634	12.3211	12.7297	13.4822	14.0104	14.622	14.7702	15.3861
	0.1	10.0089	11.0161	11.7324	12.0319	12.6967	13.163	13.6726	13.9232	14.5784
	0.3	8.7928	9.9694	10.6362	10.7009	11.3359	11.9487	12.5048	12.5833	13.2383
25	0.01	12.538	14.1579	13.9442	15.4707	14.6032	15.3676	16.3641	15.8391	16.8416
	0.03	11.4527	12.9447	13.1226	13.9027	13.7177	14.4252	15.2443	15.0588	15.8216
	0.05	10.9144	12.1809	12.6454	13.0417	13.2088	13.9915	14.6391	14.4226	15.0997
	0.1	10.0932	11.2212	11.8011	12.1558	12.5166	13.3305	13.8949	13.7394	14.5305
	0.3	8.7971	10.0375	10.6998	10.8905	11.4639	12.0895	12.6786	12.6866	13.357
30	0.01	12.8515	13.8012	14.8586	14.924	15.7008	16.4831	16.7455	17.1096	17.8048
	0.03	11.9421	13.2739	13.9407	13.8769	14.5315	15.4392	15.7004	15.9639	16.5448
	0.05	11.3231	12.6431	13.3895	13.2301	14.0336	14.8836	15.251	15.3902	16.2116
	0.1	10.6087	11.9093	12.6341	12.6507	13.3812	14.0299	14.5131	14.5987	15.3189
	0.3	9.425	10.6465	11.448	11.4694	12.1808	12.767	13.4287	13.497	14.1123
35	0.01	13.5787	14.6313	15.3557	15.8021	16.0497	16.9233	17.7182	17.911	18.6844
	0.03	12.7614	13.5108	14.4682	14.4577	14.9572	15.9731	16.4772	16.4211	17.2805
	0.05	12.1853	13.0084	13.7944	14.0042	14.4604	15.5301	15.9347	16.0617	16.7224
	0.1	11.4648	12.2917	13.0518	13.3029	13.7996	14.8474	15.213	15.3504	15.933
	0.3	10.0843	11.1736	11.9825	12.1837	12.6628	13.4633	13.9854	14.1603	14.692
40	0.01	14.1314	15.8292	16.302	16.2929	16.6779	17.1978	17.7354	17.8527	18.9472
	0.03	13.1939	14.498	15.2044	15.3041	15.6945	16.3618	17.1637	16.8675	17.6094
	0.05	12.5387	13.9559	14.4026	14.7364	15.0809	15.803	16.4463	16.469	17.2435
	0.1	11.7386	13.2732	13.7072	13.999	14.5619	15.0532	15.7521	15.7859	16.5792
	0.3	10.5784	11.8494	12.5495	12.704	13.4996	13.9472	14.6349	14.711	15.3207
45	0.01	14.5517	15.7214	16.2023	16.2815	16.7561	17.6413	18.4154	18.9821	19.0964
	0.03	13.4269	14.4359	15.3538	15.2994	15.9227	16.4005	16.8834	17.0773	17.9467
	0.05	12.9192	13.8889	14.5227	14.7263	15.3544	15.8658	16.4367	16.6616	17.2855
	0.1	12.009	13.2225	13.763	14.0352	14.7032	15.1809	15.7886	16.0087	16.6936
	0.3	10.6794	11.985	12.658	12.9085	13.3957	14.0863	14.5973	14.8029	15.4519
50	0.01	15.3759	16.258	17.0594	17.3844	17.9123	18.4991	18.335	18.4317	19.2196
	0.03	14.4233	15.1376	15.5834	15.7545	16.7638	17.1262	17.5674	17.7757	18.3132
	0.05	13.8042	14.6433	14.8799	15.1833	16.1443	16.6924	17.1546	17.335	17.7472
	0.1	13.0578	13.7686	14.2878	14.4291	15.3302	16.0272	16.4523	16.7006	17.0837
	0.3	11.4711	12.5354	13.2968	13.3107	14.0838	14.7209	15.3474	15.3893	16.0547

Table 1: The critical values for $\log(R_{n_1,n_2})$ with $\delta = 0.10$ for the two sample comparison with various sample sizes n_1 and n_2 at significance level α.

can be used in conjunction with the object (statistic) in order to make confidence sets a straightforward and automated task.

The proposed function provides the test statistic and p value, where the user can specify an option for the p value to be obtained from a Monte Carlo simulation or via interpolation from stored tables. A complementary function is also included in this package to compute the cut-off value for the appropriate tests of normality and uniformity.

To perform the goodness of fit function, we call the `dbEmpLikeGOF` function:

```
dbEmpLikeGOF(x=data, y=na, testcall="uniform", pvl.Table=FALSE, num.MC=1000)
```

where `data` represents a vector of data and the `testcall` option allows the user to perform the goodness-of-fit test for uniformity (`uniform`) or normality (`normal`). The `pvl.Table` option when set to `TRUE` employs a stored table of p values to approximate the p value for the given situation, when set to `FALSE`, a Monte-Carlo simulation scheme is employed to estimate the p value. The number of simulations in the Monte-Carlo scheme can be controlled using the `num.MC` option.

In the event that the user specifies both `x` and `y` in `dbEmpLikeGOF` the two sample distribution equality hypothesis in (15) is performed using the logarithm of the statistic in (18).

Further input options for `dbEmpLikeGOF` include specifying δ (`delta`) in (11) and δ (`delta.equality`) in (18). We recommend using the default settings and note that these procedures are fairly robust to the specification of δ.

In certain situations the user may simply be interested in obtaining the cut-off value for a given test and sample size. The function `returnCutoff` is designed to return the cut-off value for the specified goodness-of-fit test at a given α significance level. For example, the following code:

```
returnCutoff(samplesize, testcall="uniform", targetalpha=.05, pvl.Table=F,
             num.MC=200)
```

will return the Monte Carlo based test statistic cutoff for determining significance at level 0.05 for the null hypothesis in (8) with decision rule in (12).

The required input for `returnCutoff` requires the user to specify the sample size (`samplesize`) and `targetalpha` represents the significance level of the test. If the user specifies `samplesize` as a two element vector, then it is assumed that the user is specifying the two sample sizes for the distribution equality test. Note, `num.MC` represents the number of Monte-Carlo simulations performed to estimate the cut-off value. Similar to the `dbEmpLikeGOF`, there is an option to use stored tables to obtain the cutoff rather than Monte-Carlo simulations. The logical variable `pvl.Table` when true will determine the cut-off from interpolation based on stored tables. Importantly, note that the cutoff values for the test statistics in (5), (11), and (18) are returned on the logarithm scale with base e.

Using the methodology developed by North, Curtis, and Sham (2003), for each test statistic, T_{obs}, the Monte Carlo p value is computed according to the equation below:

$$p \text{ value} = \frac{1 + \sum_{j=1}^{M} I\left(T(x_1, \ldots, x_n) > T_{obs}\right)}{M + 1} \tag{20}$$

where M represents the number of simulations and $T(x_1, \ldots, x_n)$ is the statistic from the simulated data $(x_1, \ldots, x_n)$ and T_{obs} is the observed statistic.

2.1. Availability

The **dbEmpLikeGOF** package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/` and also available for download from the author's department webpage (`http://sphhp.buffalo.edu/biostat/research/software/dbEmpLikeGOF/index.php`).

3. Examples

This section provides examples of using **dbEmpLikeGOF** with the corresponding R code. Using several publicly available datasets and a novel dataset we compare our results with results from other goodness-of-fit tests including Shapiro-Wilks (SW), Kolmogorov-Smirnov (KS), Wilcoxon rank sum (WRS), and Lilliefors (L) tests. The SW test for normality was implemented using the R function shapiro.test. The one and two sample KS tests were implemented using the R function ks.test. The Lilliefors test introduced in Lilliefors (1967) is an adaptation of the Kolmogorov-Smirnov test for normality. We have included the Lilliefors test for normality as implemented in the R package `nortest` in our simulations (Gross 2006). The two sample WRS test was implemented using the R function wilcox.test (R Development Core Team 2009).

Note that Monte Carlo studies presented in Vexler and Gurevich (2010a) and Gurevich and Vexler (2011) showed various situations when the density based EL test clearly outperformed the classical procedures.

3.1. Real data examples

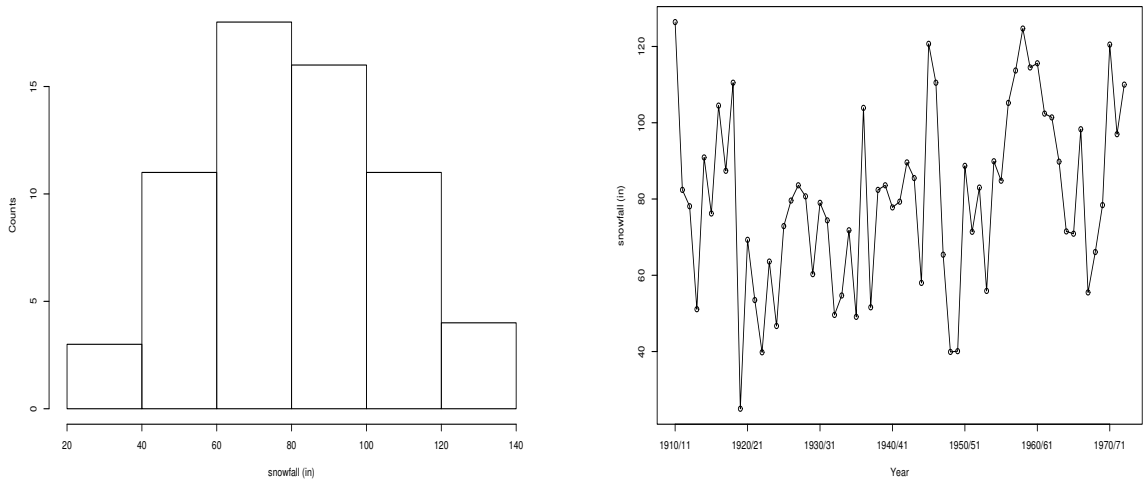
Snowfall dataset

We consider the 63 observations of the annual snowfall amounts in Buffalo, New York as observed from 1910/11 to 1972/73 (data in Table 2 and Figure 3); see, for example, Parzen (1979). We perform the proposed test for (1) with the statistic in (5). We obtain the value of the test statistic to be 8.49 with an MC based p value of 0.3234 using the following command,

dbEmpLikeGOF(x=snow, testcall="normal", pvl.Table=FALSE, num.mc=5000),

where snow represents the vector of annual snowfall amounts. Note, when using a KS test to examine the same hypothesis for the snowfall dataset we obtain a p value of 0.9851 and a SW p value of 0.5591. Thus, we conclude that there is not significant evidence to conclude that the snowfall data is inconsistent with a normal distribution.

To examine the robustness of our tests, we employed a resampling technique where we randomly removed 10, 20 and 50 percent of the data and examined the significance of the test statistics derived from the remaining dataset. For each test, we repeated this technique 2000 times where the results are summarized in Table 3. When randomly removing 10, 20, and 50 percent of the data, we obtained a significant density based EL test statistic in 3, 5.8 and 6.6



(A) (B)

Figure 3: (A) Histogram of snowfall data in Buffalo, NY from 1910/11 to 1972/73. (B) Snowfall data displayed as a time series. Using the EL test for normality, we conclude that the distribution for the data is consistent with a normal distribution (p value=.3234).

percent of the simulations, respectively. From the work in Parzen (1979), it is suggested that the Snowfall dataset follows a normal distribution. Table 4 displays the average p value for each of the tests when randomly removing the data. Ultimately, the results from randomly removing a percentage of observations demonstrates the robustness of the proposed test in controlling the type I error.

To study the power of the EL statistic, we examine four snowfall datasets where each dataset is obtained by randomly removing 50 percent of the snowfall data. These datasets represent examples where the EL based test is significant (p value < 0.05), while the KS and SW tests are not significant (p values > 0.05). These examples are summarized by displaying the kernel density estimates and the hypothesized distributions as shown in Figure 4. From the examples in Figure 4, there is the potential for the EL tests to be more powerful than KS and SW tests.

Birth dataset

As another example of `dbEmpLikeGOF`, we examine a baby boom dataset summarizing the time of birth, sex, and birth weight for 44 babies born in one 24-hour period at a hospital in Brisbane, Australia. These data appeared in an article entitled "Babies by the Dozen for Christmas: 24-Hour Baby Boom" in the newspaper The Sunday Mail on December 21, 1997. According to the article, a record 44 babies were born in one 24-hour period at the Mater Mothers' Hospital, Brisbane, Australia, on December 18, 1997. The article listed the time of birth, the sex, and the weight in grams for each of the 44 babies where the full dataset can be found at Dunn (1999). We examine whether an exponential distribution can be used to model the times between births. From the work in Dunn (1999), it is suggested that this data is exponentially distributed. The data summarizing the time between births is shown in Table 5.

Buffalo snowfall dataset ($n = 63$)								
126.4	82.4	78.1	51.1	90.9	76.2	104.5	87.4	110.5
25.0	69.3	53.5	39.8	63.6	46.7	72.9	79.6	83.6
80.7	60.3	79.0	74.4	49.6	54.7	71.8	49.1	103.9
51.6	82.4	83.6	77.8	79.3	89.6	85.5	58.0	120.7
110.5	65.4	39.9	40.1	88.7	71.4	83.0	55.9	89.9
84.8	105.2	113.7	124.7	114.5	115.6	102.4	101.4	89.8
71.5	70.9	98.3	55.5	66.1	78.4	120.5	97.0	110.0

Table 2: The amount of snowfall in Buffalo, New York, for each of 63 winters from 1910/11 to 1972/73. See Parzen (1979) for more details.

Dataset	Test	10% removed	20% removed	50% removed
Snowfall	EL	0.030	0.058	0.066
	KS	0.000	0.000	0.000
	SW	0.000	0.000	0.003
	L	0.000	0.000	0.006
Birth	EL	0.000	0.002	0.021
	KS	0.000	0.004	0.016

Table 3: Resampling results for Snowfall dataset and Birth dataset. With 2000 simulations, we randomly remove 10, 20, and 50 percent of the observations in the original dataset (Snowfall or Birth). In each remaining dataset, we compute the test statistic and the percentage of significant test statistics (at level 0.05) are summarized in each cell. EL refers to the density-based empirical likelihood test, KS denotes the Kolmogorov-Smirnov test, SW denotes the Shapiro-Wilks test, and L denotes the Lilliefors test.

Dataset	Test	10% removed	20% removed	50% removed
Snowfall	EL	0.2992	0.2956	0.3302
	KS	0.9584	0.9305	0.8779
	SW	0.5597	0.5446	0.5343
	L	0.7708	0.6950	0.6017
Birth	EL	0.6114	0.5601	0.5110
	KS	0.4652	0.5262	0.5582

Table 4: Resampling results for Snowfall dataset and Birth dataset. With 2000 simulations, we randomly remove 10, 20, and 50 percent of the observations in the original dataset (Snowfall or Birth). In each remaining dataset, we compute the test statistic and the mean p values are summarized in each cell. EL refers to the density-based empirical likelihood test, KS denotes the Kolmogorov-Smirnov test, SW denotes the Shapiro-Wilks test, and L denotes the Lilliefors test.

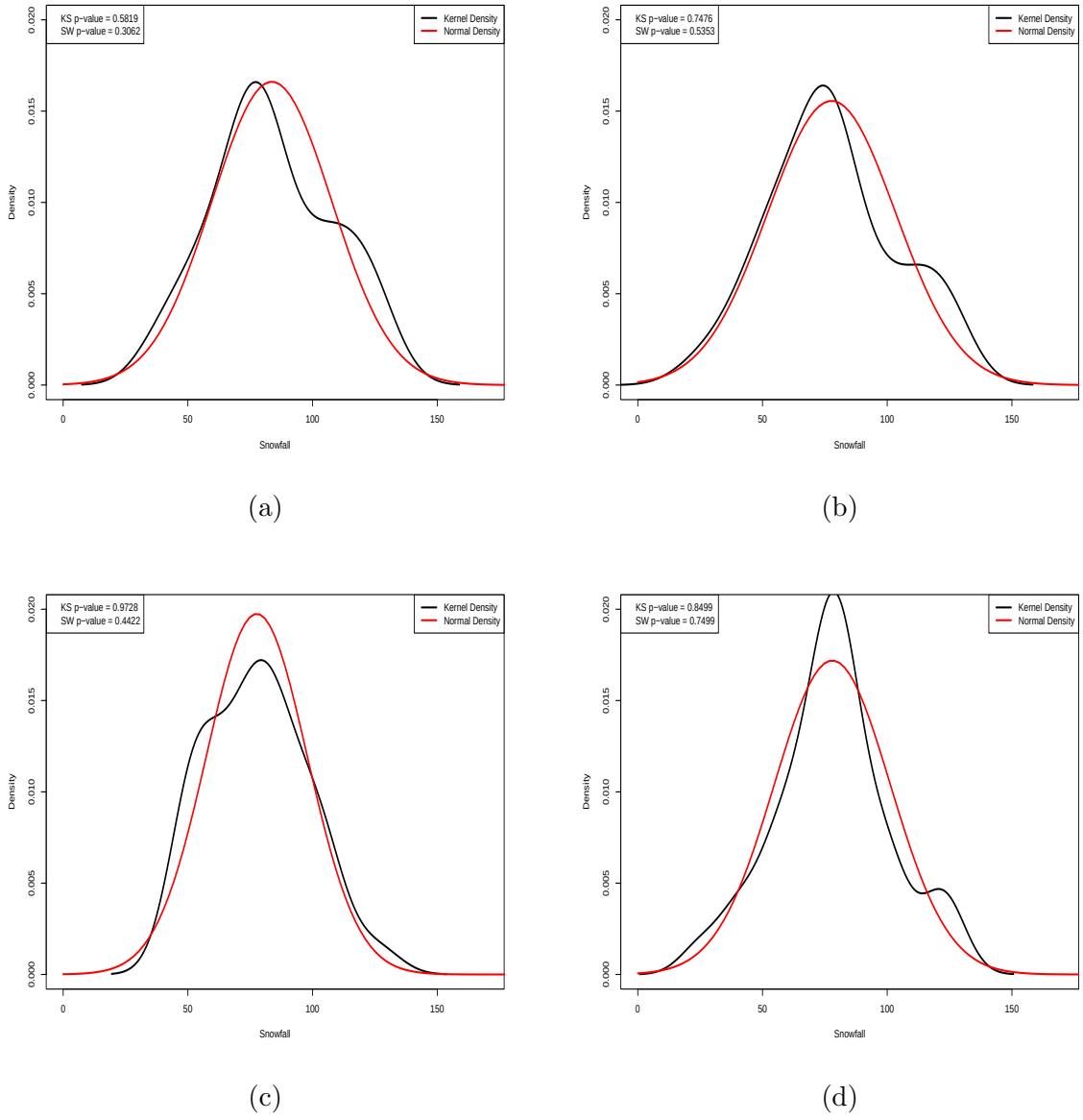


Figure 4: Snow fall dataset examples where the density-based EL statistic is significant (p value < 0.05), but the Kolmogorov-Smirnov (KS) test and Shapiro Wilks (SW) test are not significant (p values > 0.05). The normal density (red curve) is determined using the sample mean and sample standard deviation to estimate the mean and standard deviation. The black curve represents the kernel density for a randomly chosen subset from the snowfall dataset described in Section 3.1.

Using the KS test for an exponential distribution, we obtain a p value of 0.3904. We transform the data in Table 5 using the inverse exponential distribution and thus the transformed data can be examined using the EL ratio test for uniformity. The following command returns the test statistic (11) and p value,

Inter-time births ($n = 43$)		
Time between births (minutes)	Tally	Empirical probability
00-19	18	0.419
20-39	12	0.279
40-59	6	0.140
60-79	5	0.116
80+	2	0.047
Total	43	1.001

Table 5: The time between births for 44 babies born in one 24 hour period at the Mater Mothers' Hospital, Brisbane, Australia, on December 18, 1997. See Dunn (1999) for more details.

```
dbEmpLikeGOF(x=baby, testcall="uniform", pvl.Table=FALSE, num.mc=5000),
```

where baby represents the vector of transformed data. When this test is employed, we observe a MC based p value of 0.6708. Ultimately, for this data the time between births can be adequately modeled using an exponential distribution.

Similar to the snowfall dataset, we examine the robustness of our results by employing a bootstrap scheme where the bootstrap resamplings are taken when removing 10, 20, or 50 percent of the original dataset. The results are summarized in Tables 3 and 4. With 2000 simulations where we randomly remove 10, 20, and 50 percent of the observations from the original dataset, we find significant statistics in 0, .2, and 2 percent of the simulated datasets, respectively.

To examine the power of the EL statistic, we examine four birth datasets where the EL test is significant (p value < 0.05), while the KS test is not significant (see Figure 5). Figure 5 displays the data driven kernel density estimate against the hypothesized distribution. From these examples, there may be situations where the EL test for uniformity may be more powerful than the traditional KS test.

A TBARS data example

For a novel analysis using the density-based EL software, we consider data from a study evaluating biomarkers related to atherosclerotic (CHD) coronary heart disease (see Acknowledgments). A population-based sample of randomly selected residents of Erie and Niagara counties of the state of New York, U.S.A., was the focus of this investigation. The New York State Department of Motor Vehicles drivers' license rolls were utilized as the sampling frame for adults between the ages of 35 and 65; where the elderly sample (age 65-79) was randomly selected from the Health Care Financing Administration database. Participants provided a 12-hour fasting blood specimen for biochemical analysis at baseline, and a number of parameters were examined from fresh blood samples. A complete description of this dataset is available at Schisterman, Faraggi, Browne, Freudenheim, Dorn, Muti, Armstrong, Reiser, and Trevisan (2001).

A cohort of 5620 men and women were selected for the analyses yielding 1209 cases (individuals that had a heart attack) and 4411 controls (no heart attack history). In a subset of this dataset, we examine the significance of the thiobarbituric acid reactive substances (TBARS)

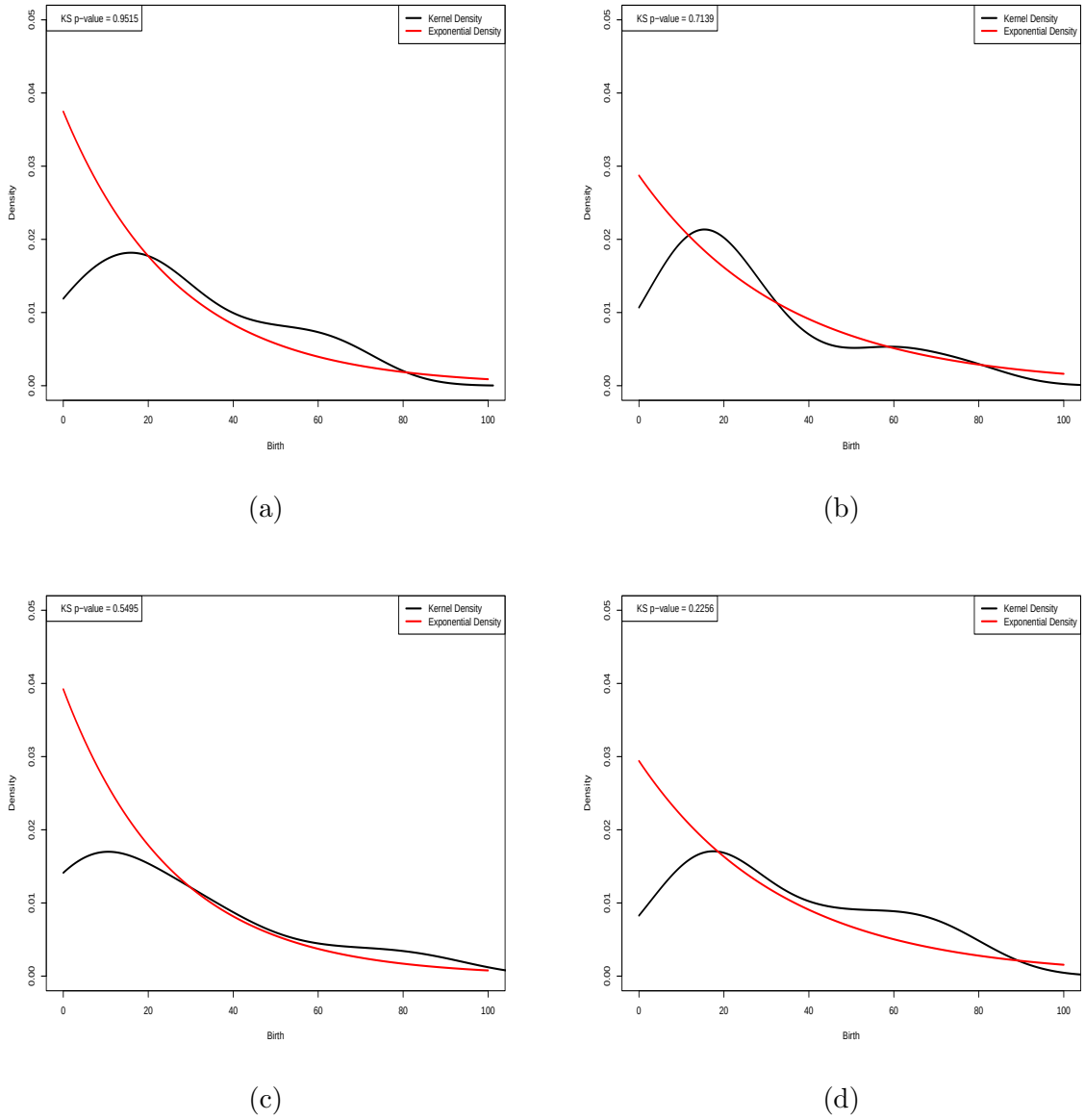


Figure 5: Birth dataset examples where the density-based EL statistic is significant (p value < 0.05), but the Kolmogorov-Smirnov (KS) test is not significant (p values > 0.05). The exponential density (red curve) has a rate parameter of the inverse of the sample mean. The black curve represents the kernel density for a randomly chosen subset from the birth dataset described in Section 3.1.

variable which is known to play a role in atherosclerotic coronary heart disease process. TBARS was measured in patient serum samples using reverse-phase high performance liquid chromatography and spectrophotometric approaches.

For the analysis of the TBARS dataset, we would like to test the claim that the TBARS distribution is different between the cohort of patients that have suffered a heart attack and

the cohort of patients that have not suffered a heart attack. If the null hypothesis is true, we expect the empirical distribution of the TBARS variable to be very similar in the two cohorts, if the null hypothesis is not true, we expect the empirical distributions to be very different (e.g. TBARS is stochastically greater in the heart attack population). A quantile-quantile (QQ) plot of this data is shown in Figure 6.

We employed a bootstrap strategy to study the TBARS variable using the statistic in (18). The strategy was based on randomly choosing 200 patients, where 100 patients had previously suffered a heart attack and 100 patients did not have a heart attack. The distribution of TBARS was examined for equality between the heart attack patient cohort and the no heart attack patient cohort. We repeated this procedure 2000 times calculating the frequency of the event of a significant statistic. Rather than obtain a p value associated with each statistic, we employed the returnCutoff command to obtain the cutoff for significance,

```
tbar.cut = returnCutoff(100, testcall="distribution.equality", targetalpha=0.05, n.mc=5000).
```

Figure 7 displays the histogram of the logarithm of the two sample test statistic calculated in (18). 5.9 percent of the bootstrap resamplings yielded a significant test statistic. These results were also compared against the two sample KS test and two sample WRS test to compare distribution equality. Using the KS test, 3.4 percent of the resamplings were significant (p value < 0.05). Using the two sample WRS test, 4.5 percent of the resamplings were significant (p value < 0.05).

Thus all of the statistical tests suggest that TBARS is not significantly different in heart attack patients. This is further confirmed when examining the mean p-values in the resampled datasets. The mean p-values are 0.5262, 0.5002, and 0.4489 for the KS, WRS, and empirical likelihood tests, respectively.

To examine the power of the two sample EL test, we focus on four examples comparing 100 patients that had previously suffered a heart attack and 100 patients that did not have a heart attack, where the EL statistic is significant, however, the KS and WRS tests are not significant. Figure 8 displays the density for each cohort when a kernel density smoother is employed with the bandwidth chosen according to Equation (3.31) in Silverman (1986). These examples highlight situations where there may be a power advantage obtained using the density-based EL statistics over traditional goodness-of-fit tests such as the two sample KS test and two sample WRS test.

Acknowledgments

The authors are grateful to Dr. Jo L. Freudenheim, chairperson of the Department of Social and Preventive Medicine at the SUNY University at Buffalo, for generously allowing us to use the TBARS dataset.

Further, the authors would also like to acknowledge and thank the Editors and referees for their helpful comments and suggestions that improved this article.

4. Conclusions

The package **dbEmpLikeGOF** provides R users with a new and powerful way to perform

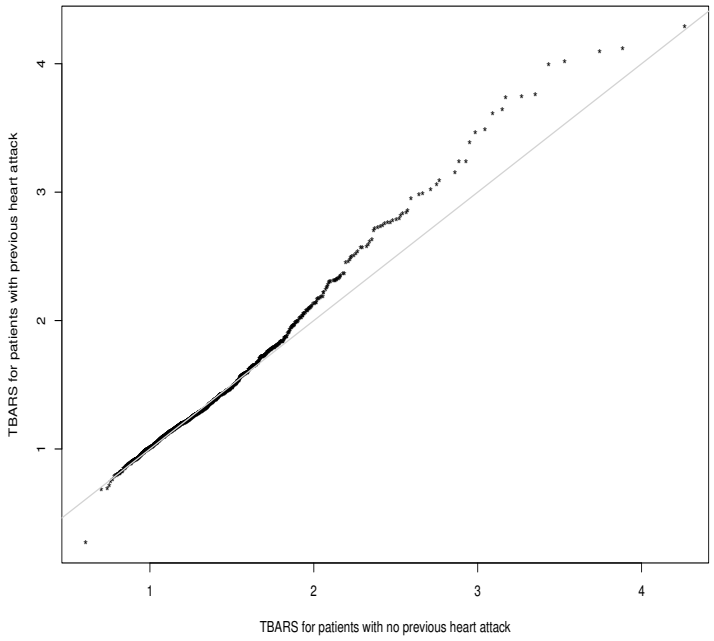


Figure 6: A quantile-quantile (QQ) plot comparing the distribution of TBARS for patients with a previous heart attack against the distribution of TBARS for patients without a previous heart attack.

goodness-of-fit tests using empirical likelihood ratios. We focus on two sample tests and tests for normality and uniformity which are common distributions to test in applied studies. Monte-Carlo methods and interpolation are used to estimate the cutoff-values and *exact* p values for the proposed tests. The proposed procedure can execute entropy based structured tests that have not been addressed in statistical software. We believe that the **dbEmpLikeGOF** package will help investigators to use density based empirical likelihood approaches for goodness-of-fit tests in practice.

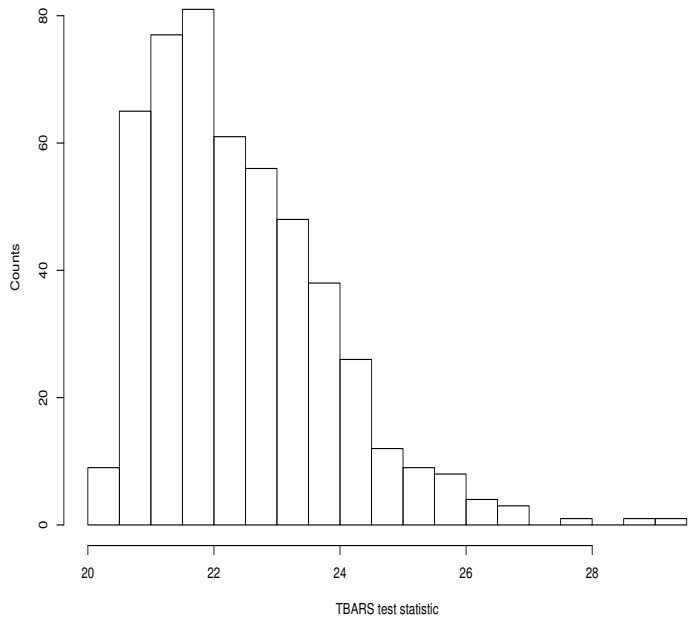


Figure 7: Histogram of the test statistic for TBARS distribution equality based on 2000 bootstrap resamplings.

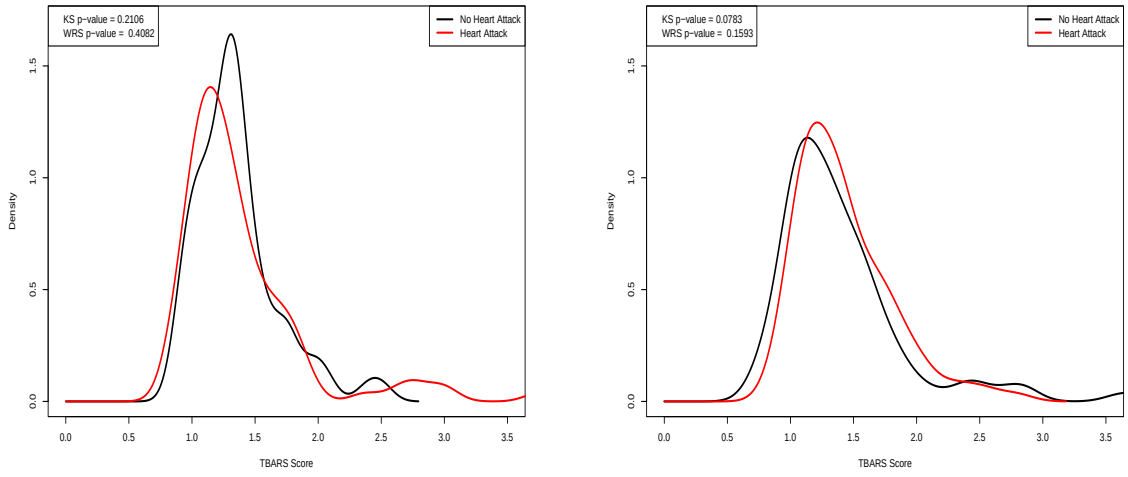


Figure 8: TBARS examples where the density-based EL distribution equality statistic is significant (p value < 0.05), but the two-sample Kolmogorov-Smirnov (KS) test and two-sample Wilcoxon rank sum (WRS) test are not significant (p values > 0.05).

References

Darling D (1957). "The Kolmogorov-Smirnov, Cramér-von Mises tests." *The Annals of Mathematical Statistics*, **28**(4), 823–838.

Dodge Y (2006). *The Oxford dictionary of statistical terms*. Oxford University Press, USA.

Dunn P (1999). "A simple data set for demonstrating common distributions." *Journal of Statistics Education*, **7**(3).

Gross J (2006). *nortest: Tests for Normality*. R package version 1.0.

Gurevich G, Vexler A (2011). "A two-sample empirical likelihood ratio test based on samples entropy." *Statistics and Computing*, pp. 1–14.

Hollander M, Wolfe D, Wolfe D (1973). *Nonparametric statistical methods*. Wiley New York.

Lilliefors H (1967). "On the Kolmogorov-Smirnov test for normality with mean and variance unknown." *Journal of the American Statistical Association*, pp. 399–402.

North B, Curtis D, Sham P (2003). "A note on the calculation of empirical P values from Monte Carlo procedures." *American journal of human genetics*, **72**(2), 498.

Owen A (2001). *Empirical likelihood*. CRC press New York. ISBN 1584880716.

Parzen E (1979). "Nonparametric statistical data modeling." *Journal of the American Statistical Association*, **74**(365), 105–121.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Royston P (1991). "Estimating departure from normality." *Statistics in Medicine*, **10**(8), 1283–1293.

Schisterman E, Faraggi D, Browne R, Freudenheim J, Dorn J, Muti P, Armstrong D, Reiser B, Trevisan M (2001). "TBARS and cardiovascular disease in a population-based sample." *European Journal of Cardiovascular Prevention & Rehabilitation*, **8**(4), 219.

Silverman B (1986). *Density Estimation for Statistics and Data Analysis*, volume 26. Chapman & Hall/CRC.

Vexler A, Gurevich G (2010a). "Density-Based Empirical Likelihood Ratio Change Point Detection Policies." *Communications in Statistics-Simulation and Computation*, **39**(9), 1709–1725.

Vexler A, Gurevich G (2010b). "Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy." *Computational Statistics & Data Analysis*, **54**(2), 531–545.

Vexler A, Yu J, Tian L, Liu S (2010). "Two-sample nonparametric likelihood inference based on incomplete data with an application to a pneumonia study." *Biometrical Journal*, **52**(3), 348–361.

Yu J, Vexler A, Tian L (2010). "Analyzing incomplete data subject to a threshold using empirical likelihood methods: An application to a pneumonia risk study in an ICU setting." *Biometrics*, **66**(1), 123–130.

Affiliation:

Jeffrey Miecznikowski
Department of Biostatistics
University at Buffalo
Kimball Tower Rm 723
3435 Main Street, Buffalo NY 14214 E-mail: jcm38@buffalo.edu
URL: http://sphhp.buffalo.edu/biostat/faculty/miecznikowski_jeff.php/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd